DOCUMENT RESUME

ED 266 439 CS 008 341

AUTHOR Glenberg, Arthur M.; Epstein, William
TITLE Inexpert Calibration of Comprehension. Program Report 86-3.
INSTITUTION Wisconsin Center for Education Research, Madison.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.; Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE 1 Mar 86
CONTRACT N0014-85-K-0063
GRANT NIE-G-84-0008
NOTE 40p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Content Area Reading; Higher Education; *Inferences; Knowledge Level; Reading Ability; *Reading Comprehension; *Reading Research; *Self Esteem; *Self Evaluation (Individuals)
IDENTIFIERS *Calibration; *Metacomprehension

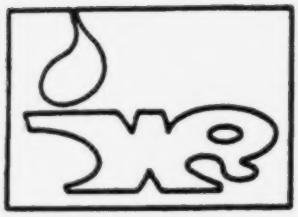
ABSTRACT

A total of 70 college students who had a wide range of coursework in physics or music theory read expositions in both domains in a study to determine the relationship between readers' self-assessments of confidence in comprehension (strength of belief) and performance on a test of comprehension. After reading each text, students provided a judgment of confidence in ability to verify inferences based on the central principle of the text. The primary dependent variable was calibration of comprehension. Findings showed that expertise in a domain was inversely related to calibration and subjects were well-calibrated across domains. The results suggested that confidence judgments were based on self-classification as expert or nonexpert in the domain of the text, rather than an assessment of the degree to which the text was comprehended. (Author/HOD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED266439



Program Report 86-3
March 1, 1986

Inexpert Calibration of Comprehension

by Arthur M. Glenberg
and William Epstein

This work was supported by National Institute of Education Grant NIE-G-84-0008 and by Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under Contract No. N0014-85-K-0063, Contract Authority Identification Number, NR 702-012. Approved for public release; distribution unlimited. Reproduction in whole or part is permitted for any purpose of the United States Government.

Wisconsin Center for Education Research
School of Education, University of Wisconsin-Madison

CS008341

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) WCER Program Report 86-3	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Wisconsin Center for Education Research 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Personnel and Training Research Programs Office of Naval Research (Code 1142PT)
6c. ADDRESS (City, State, and ZIP Code) 1025 W. Johnson Street Madison, WI 53706	7b. ADDRESS (City, State, and ZIP Code) 800 North Quincy Street Arlington, VA 22217-5000
8a. NAME OF FUNDING / SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N0014-85-K-0644
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
61153N	RR04206	OC	NR702-012

11. TITLE (Include Security Classification)
Inexpert Calibration of Comprehension (Unclassified)

12. PERSONAL AUTHOR(S)
Glenberg, Arthur M., & Epstein, William

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical Report	FROM 9/85 TO 12/85	86-3-1	46

16. SUPPLEMENTARY NOTATION
Under review

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	comprehension, meta-comprehension, expert knowledge
05	10		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Students with a wide range of coursework in physics or music theory read expositions in both domains. After reading, for each text students provided a judgment of confidence in ability to verify inferences based on the central principle of the text. The primary dependent variable was calibration of comprehension, the degree of association between confidence and performance in the inference test. Two results of most interest were (a) expertise in a domain was inversely related to calibration and (b) subjects were well-calibrated across domains. Both of these results can be accommodated by a self-classification strategy: Confidence judgments are based on self-classification as expert or non-expert in the domain of the text, rather than an assessment of the degree to which the text was comprehended. Because self-classifications are not well differentiated within a domain, application of the strategy by experts produces poor calibration within a domain. Nonetheless, because self-classification is generally consistent with performance across domains, application of the strategy produces calibration across domains.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION	
☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	Unclassified	
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. Michael Shafto	22b. TELEPHONE (Include Area Code) 202-696-4596	22c. OFFICE SYMBOL ONR 1142PT

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

Program Report 86-3

Inexpert Calibration of Comprehension

Arthur M. Glenberg and William Epstein

University of Wisconsin

Wisconsin Center for Education Research

School of Education

University of Wisconsin-Madison

March 1, 1986

This work was supported by National Institute of Education Grant NIE-G-84-0008 and by Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under Contract No. N0014-85-K-0063, Contract Authority Identification Number, NR 702-012. Approved for public release; distribution unlimited. Reproduction in whole or part is permitted for any purpose of the United States Government.

WISCONSIN CENTER FOR EDUCATION RESEARCH

MISSION STATEMENT

The mission of the Wisconsin Center for Education Research is to improve the quality of American Education for all students. Our goal is that future generations achieve the degree of knowledge, tolerance, sensitivity, and complex thinking skills necessary to ensure a productive and enlightened democratic society. We are willing to explore solutions to major problems, recognizing that radical change may be necessary to meet our goal.

Our approach is interdisciplinary because the problems of education in the United States go far beyond pedagogy. We therefore draw on the knowledge of scholars in psychology, sociology, history, economics, philosophy, and law as well as experts in teacher education, curriculum, and administration in order to arrive at a deeper understanding of schooling.

Work of the Center clusters in four broad areas:

- Learning and Development focuses on individuals, in particular on their variability in basic learning and development processes.
- Classroom Processes seeks to adapt psychological constructs to the improvement of classroom learning and instruction.
- School Processes focuses on schoolwide issues and variables, seeking to identify administrative and organizational practices that are particularly effective.
- Social Policy is directed toward delineating the conditions under which social policy is likely to succeed, the ends to which it is suited, and the constraints which it faces.

The Wisconsin Center for Education Research is a noninstructional unit of the University of Wisconsin-Madison School of Education. The Center is supported primarily with funds from the Office of Educational Research and Improvement/Department of Education, the National Science Foundation, and other governmental and non-governmental sources in the U.S.

Abstract

Students with a wide range of coursework in physics or music theory read expositions in both domains. After reading, for each text students provided a judgment of confidence in ability to verify inferences based on the central principle of the text. The primary dependent variable was calibration of comprehension, the degree of association between confidence and performance on the inference test. Two results of most interest were (a) expertise in a domain was inversely related to calibration and (b) subjects were well-calibrated across domains. Both of these results can be accommodated by a self-classification strategy: Confidence judgments are based on self-classification as expert or non-expert in the domain of the text, rather than an assessment of the degree to which the text was comprehended. Because self-classifications are not well differentiated within a domain, application of the strategy by experts produces poor calibration within a domain. Nonetheless, because self-classification is generally consistent with performance across domains, application of the strategy produces calibration across domains.

A reader's self-assessment of comprehension often has significant consequences for the reader's action. When reading under time constraints, the reader's belief that comprehension has been achieved will encourage the reader to terminate further processing of the text. When reading in preparation for testing, the belief that comprehension has been attained will lead the reader to declare his readiness for testing. Given these and other implications for action, it is sensible to inquire whether readers' beliefs are regularly valid. Taking as our measure, the relationship between the readers' self-assessments of confidence in comprehension (strength of belief) and performance on a test of comprehension, we have repeatedly found that readers' beliefs typically are off the mark. Readers are very poorly calibrated: confidence in comprehension (belief) does not predict performance.

Glenberg and Epstein (1985) measured calibration by having subjects read 15 short expositions on a variety of topics. Subjects also provided an assessment of their confidence in ability to use a principle from the text (provided at the time of the confidence assessment) to judge whether or not an inference was correct. Finally, subjects attempted to decide if an inference using the principle was or was not valid. One measure of calibration of comprehension is the point biserial correlation between the confidence assessments and performance on the inference test. In none of three experiments reported by Glenberg and Epstein was this correlation significantly different from zero.

In subsequent unpublished experiments deploying a variety of performance measures and a diverse set of measures of calibration, the finding of zero or marginal calibration has recurred. This result is disconcerting because it appears to identify an important obstacle in learning from text. The result

also does not conform to our personal experience. In our experience in learning from text, calibration of comprehension seems reasonably good.

Upon more detailed scrutiny of our experience, our initial impression that, in general, we were calibrated had to be qualified. Our impression may have been much affected by the availability heuristic. In assessing the degree of calibration that we exhibited we relied heavily on the most readily available instances, and as a matter of course, these were instances involving texts in our personal domains of expertise. By contrast, in our experiments, the texts were by design a varied set that probably touched only peripherally on readers' special fields of competence. These considerations led to the current experiment to test the relationship between calibration and expertise.

Everyday observation suggests that experts may be well-calibrated. These observations are probably confounded with the domain of reading, however. That is, the expert knows that he is competent in the domain of expertise and that he is less competent in other domains. Thus by using base rates the expert can accurately predict better performance in the domain of expertise than in alternative domains. Nonetheless, this ability to predict relative performance across domains does not imply that the expert is well calibrated within a domain.

In fact, a sampling of the literature indicates that relative expertise does not confer an ability to predict performance within the domain. Oskamp (1965) has reported that trained clinical psychologists are greatly overconfident in their predictions derived from reading case studies. Similarly, Hock (1985) found that students in a master's in business administration program were overconfident in their predictions of their future success in developing employment opportunities. Bradley (1981) had

undergraduates rank their knowledge in twelve domains. He then administered a short test on content from each domain and had subjects rate confidence in each answer. Performance on the test was positively related to the knowledge rankings. However, confidence in incorrect answers also increased with the knowlege ranking. The "experts" were less likely (or willing) to admit ignorance.

We recruited subjects who had a minimum of two college-level physics courses or two college-level music courses (excluding performance courses such as marching band). Within each of these groups subjects had a wide range of formal coursework and non-academic experience. We choose these two domains because, the knowledge acquired within the domains have little overlap. Also, Birkmire (1982) has found that music students reading in the domain of music were more sensitive to structurally important components of the text than when reading in the domain of physics. Physics students showed the converse effect.

Our stimulus materials were prepared by two graduate students: a graduate student in physics composed 16 expositions on various topics in physics; a graduate student in music theory composed 16 expositions on various topics in music. Each of the subjects read all of these texts, eight physics texts and eight music texts on each of two days. At the end of each day's session, the subject rated confidence in ability to correctly answer inferences for each text and was given the inference verification test. (Glenberg and Epstein (1985) demonstrated that delaying the confidence assessment and the test until the end of a session does not change calibration.)

The expertise hypothesis predicts that physics students will be better calibrated for the physics texts than for the music texts, and that music students will show the opposite pattern. On the other hand, expertise may only

confer the ability to predict better performance in the domain of expertise than in an alternative domain. In this case, (a) experts will be poorly calibrated in both domains, but (b) calibration computed across domains will be greater than zero.

The experiment was also designed to assess a number of other questions. First, Glenberg and Epstein (1985) found that, although the average measure of calibration was not significantly different from zero, there was large variation in the point biserial correlations. Having subjects read texts on two days allowed us to determine if this variability is due to random error or stable individual differences.

In addition to obtaining information from subjects regarding their experiences in the domains of physics and music, each subject was assessed on the dualism scale (Ryan, 1984). A dualist has relatively immature epsitemological standards, believing that truth is absolute in most if not all domains. A relativist believes that truth is determined by the context, that propositions are true or false within a particular frame of reference. Ryan demonstrated that relativists engage in more sophisticated comprehension monitoring than do dualists. Thus if there are stable individual differences in calibration of comprehension, the tendency toward dualism may well predict those differences.

The experiment was also designed to test the generality of two other findings reported by Glenberg and Epstein (1985). In their third experiment, subjects provided three responses after answering the inference question for each text. First, the subject was asked to rate confidence in the correctness of the answer to the inference question. The correlation of this confidence rating and performance on the test is called calibration of performance. In

contrast to initial calibration, calibration of performance was significantly greater than zero. This finding is consonant with Lichtenstein, Fischhoff, and Phillips's (1982) results that accuracy of postdictions are significantly better than chance (although generally exhibiting overconfidence).

After rating confidence in performance, subjects in Glenberg and Epstein's third experiment provided another assessment of confidence in ability to judge inferences on an upcoming test. Then a second inference test was given. The correlation between this second prediction and performance on the second test is called recalibration. In Glenberg and Epstein's third experiment, recalibration was significantly greater than zero. Glenberg and Epstein proposed that the experience gained from answering the first inference question (e.g., ease of retrieval of relevant propositions, amount of time required to check the inference) provided valid cues to the degree of comprehension, and that these cues could be used to predict future performance. A similar hypothesis has been offered to explain the relationship between accuracy and confidence in eye-witness identification. Kassin (1985) found that subjects in the eye-witness identification task are generally poorly calibrated. Having subjects attend to the experience of making a judgement results in significant improvements in calibration.

The current experiment includes the measurements needed to compute both calibration of performance and recalibration. Either of these measures may be related to expertise in a domain of knowledge.

Method

Subjects

A total of 70 subjects was recruited from the University of Wisconsin-Madison community. A variety of recruitment procedures were used

including posters advertising the experiment, mailings to students meeting the minimum coursework requirements, and solicitation in upper-level classes. The minimum coursework requirement was completion of two university-level courses in either physics or music theory. Upon completing the experiment, subjects completed a questionnaire requiring a listing of the university-level music and physics courses completed, as well as listing other experiences either in music (e.g., lessons on an instrument) or physics (working as a laboratory assistant). These experiences were coded using a scale of 0 (no experience) to 3 (experience at a professional level such as giving music lessons). Descriptive statistics are given in Table 1.

Insert Table 1 about here

Since there were subjects who had relevant experience in both music and physics, we did not attempt to classify subjects into mutually exclusive categories. Instead, background knowledge was coded using four variables, number of music courses, music experience, number of physics courses, and physics experence. These four variables were then entered, as a set, into a hierarchical multiple regression analysis to determine the effect of background knowledge on calibration.

The questionnaire also contained a seven-item scale for measuring dualism (Ryan, 1984). Subjects rated the relative frequency (1= rarely, 5= almost always) of experiencing thoughts such as "If professors would stick more to the facts and do less theorizing one could get more out of college." The higher the average rating, the greater the tendency toward dualism. Data from this scale are also given in Table 1.

Subjects were paid $8.00 for participating in the experiment.

Materials

Each text was one paragraph long and was written to illustrate or explicate a central principle that was stated explicitly in the text. An example is presented in the appendix with the central principle highlighted. The principle was not highlighted for the subjects. Two pairs of inference questions were written for each text. Each of these questions stated an inference that the subject was to judge as true or false. One member of each pair was a true inference, the other member of each pair was a false inference. Accurate performance on the inference tests required knowledge of the central principle. Examples of the inference tests are provided in the appendix.

The texts were arranged in two booklets with 16 texts in each. One booklet was used for the first session, and one booklet was used for the second. Within each booklet there were eight music texts alternating with eight physics texts. The order of the texts was counterbalanced over subjects.

Following the texts in each booklet were 16 sets of five probes. Each set corresponded to one of the texts, and the sets were in the same order as the texts. The confidence probe (probe 1) gave the title of the text and required the subject to indicate confidence in ability to judge the correctness of an inference regarding -------- . The blank was filled with a reference to the central principle (see the appendix for examples). Subjects responded by circling a confidence rating of 1 (very low) to 6 (very high).

The inference test (probe 2) was on the following page (headed by the title of the relevant text). Subjects judged the correctness of the inference by circling a T (true) or F (false). The confidence in performance scale (probe 3) was on the same page. Subjects were asked to rate their confidence that they

had answered the inference test correctly (using a number from 1 to 6). The recalibration confidence scale (probe 4) was also on this page. Subjects indicated confidence in ability to answer another inference regarding the central principle. Once again, confidence was indicated by circling a number from 1 to 6.

The following page presented the second inference test (the fifth probe). This page was also headed by the title of the text. Again, subjects responded by circling T or F.

Procedure

Subjects were tested in small groups. The instructions explained that the aim of the experiment was to investigate how students assess comprehension. They were told that they could read the passages at their own pace, and re-reading of a passage was allowed. However, once any page was turned, it could not be turned back. Further instruction regarding how to answer the five probes was also provided.

On the first day, the experiment was adjourned after subjects had read and completed the 16 sets of probes. The second session was scheduled for 1 to 7 days later. At the end of the second session the subjects completed two questionnaires. For the first, subjects were asked to rate the familiarity of each of the 32 texts on a scale of 1 to 6. Subjects were provided with copies of the texts while producing the ratings. The second questionnaire was the survey on domain-specific experiences and dualism.

Results

The basic strategy of data analysis was to use hierarchical multiple regression techniques to perform an analysis of variance (Cohen & Cohen, 1977). Two groups of analyses were performed. In the initial analyses the

between-subjects variables were dualism entered into the regression first, followed by the four background knowledge variables entered as a set with four degrees of freedom. The protected-*t* procedure was used; the significance of individual components of the background knowledge set were only examined when the omnibus *F* was significant. The within-subjects variables were type of text (music or physics) and the interaction of type of text and background knowledge. The protected-*t* procedure was also used to examine components of this interaction. The interaction of dualism and type of text was not examined. The *MSE* terms were computed by dividing the proportion of (between-subject or within-subject) variance not accounted for by any of the independent variables by the degrees of freedom.

The second set of analyses was motivated by two concerns. First, the dualism variable accounted for little variance and thus tended to waste degrees of freedom. Second, there were significant positive correlations between music experience and music courses variables (.62) and between physics experience and physics courses (.47). These correlations can distort the significance levels of the the individual variables when they are entered as a set (the problem of collinearity, Cohen & Cohen, 1975). For these reasons, the second set of analyses omitted the dualism, music experience, physics experience variables. Fortunately, the second set of anaylses produced a very similar pattern of significant results as the first set of analyses. Because the second analyses are simpler, they will be the main focus of the results section. Reference to the first analyses will only be made when there is a significant discrepancy between the two.

The measurement of calibration requires variability in both the use of the confidence scale and in performance on the inference test. Because some

subjects used the same confidence judgement or answered all of the inference questions correctly, they were excluded from some of the analyses. Consequently, the number of subjects contributing to each analysis differed. This number is indicated at the beginning of each of the sections dealing with separate analyses.

Initial calibration and its components

Confidence (probe 1), n = 61. The mean confidence on the music texts (with standard deviation in parentheses) was 4.69 (.99), and the mean confidence on the physics texts was 4.73 (.94). These means were not significantly different. There was one significant effect in the analysis of variance, type of text interacted with background knowledge, $F(4, 116) = 79.34$, MSE = .0024. Both of the background knowledge variables, number of music courses and number of physics courses, were significant contributors to this interaction.

Insert Table 2 about here

The regression coefficients are given in Table 2. These coefficients indicate the average change in the dependent variable (in this case, confidence) for each unit change in the independent variable.

The coefficients in Table 2 indicate a reasonable pattern of relationships between the independent variables and confidence. Confidence in music texts increases with the number of music courses, and the increase for music texts is significantly greater than the increase for the physics texts. Also, confidence in physics texts increases with number of physics courses, and that increase is significantly greater for the physics texts than for the music texts.

These results provide a manipulation check on the construction and classification of the texts, and the validity of the the background knowledge variables. That is, the interaction between text type and confidence is just what would be expected if our subjects did indeed differ in expertise in the two fields, and the texts tapped that difference.

Proportion correct on the first inference test (probe 2), n = 61. Mean proportion correct was .72 (.12) on the music texts and .79 (.12) on the physics texts, a significant difference, $F(4, 116) = 38.39$, MSE = .0021. The set of background knowledge variables also accounted for a significant part of the variance, $F(2, 58) = 8.48$, MSE = .0133. Only the physics courses variable was significant by the protected-t procedure. Each additional physics course was associated with a .0217 increase in proportion correct (averaged over both types of text).

In the first analyses of proportion correct, a significant main effect was found for dualism, $F(1, 55) = 4.54$, MSE = .0129. Each unit increment on the dualism scale was associated with a .0268 reduction in proportion correct.

There was also a significant interaction between type of text and background knowledge, $F(2,116) = 19.42$, MSE = .0021. The regression coefficents for this interaction are given in Table 2. The major component carrying the interaction was number of music courses. Proportion correct on the music texts increased with increases in music courses, whereas proportion correct on the physics texts was essentially unrelated to music courses. The opposite pattern was found for the physics courses variable (although not significant): Proportion correct on the physics tests increased more with physics experience than did proportion correct on the music texts. The failure to reach significance may in part reflect the problem of collinearity. The two variables are significantly, although negatively, correlated (-.44).

Calibration of comprehension, n = 50. Calibration is measured by the degree of association between confidence and performance on the inference test. One such measure is the point-biserial correlation. Unfortunately, this measure has a number of undesireable properties, including that the maximum value depends on the proportion correct. Nelson (1984) suggests that the Goodman-Kruskal gamma (G) is the most appropriate index of association for measuring metacognitive performance under the conditions instantiated in this experiment. Gamma ranges from -1 to 1, with 0 indicating no relationship. It has a direct interpretation in terms of the difference between two probabilities. Consider all pairs of texts that for a given subject, differ on both confidence and performance on the inference test. Gamma is the difference between the probability that the text with the greater confidence has the better performance and the probability that the text with the greater confidence has the lower performance.

For each subject, G was computed separately for the music texts and for the physics texts. The means were .06 (.53) for the music texts and .02 (.62) for the physics texts. Neither of these means was significantly different from zero, nor were they different from one another. Although none of the main effects were significant, there was a significant interaction between type of text and background knowledge, $F(2, 94) = 7.99$, MSE = .0044. The regression coefficients for this interaction are given in Table 2. The significant component of the interaction was the interaction of text type and number of physics courses. An increase in number of physics courses tended to decrease G for the physics texts, but had essentially no relationship to G for the music texts.

The finding of no overall calibration of comprehension replicates our previous results (Glenberg & Epstein, 1985). The new information provided by

this experiment concerns the relationship between level of knowledge in a domain and calibration in that domain. Under these experimental conditions that relationship is negative. Note that for the physics texts, subjects with no physics courses and the average number of music courses (2.76) are predicted by the regression equation to be fairly well calibrated, G = .3152. However, the predicted G drops to .0170 for subjects with the average number of both music and physics courses. This new result is discussed further in Discussion section.

Calibration of Performance

Insert Table 3 about here

Confidence in performance (probe 3), n = 61. After answering an inference question, subjects rated confidence in his or her answer to the inference question. The mean confidence ratings were 4.76 (.73) and 4.99 (.67) for the music and physics texts, respectively. These means were significantly different, $F(1, 116) = 12.22$, MSE = .0021. There was also a significant interaction between type of text and background knowledge, $F(2, 116) = 59.59$, MSE = .0021. Each of the background knowledge variables contributed to this interaction, ts > 3.65.

The regression coefficients are given in Table 3. Note that the pattern of the coefficients differs for confidence (probe 1, Table 2) and confidence in performance (probe 3, Table 3). That is, for both variables, the difference between the coefficients for music texts and physics texts is smaller in Table 3 than in Table 2. We will use this difference to argue (in the Discussion section) that subjects used different strategies to produce the two confidence ratings.

Calibration of performace (probes 2 and 3), n = 55. Is there a significant relationship (G) between confidence in performance and actual performance? In short, the answer is yes. The average performance G for the music texts was .42 (.43) and the average for the physics texts was .36 (.55). Both of these Gs are significantly greater than zero, and they are sizeable on an absolute scale. Remember that G is a difference in probabilities: An average G of .39 means that for texts that differ in confidence and whether or not they are correct on the inference test, the probability that the text with the greater confidence is correct is .39 greater than the probability that the text with the lower confidence is correct.

Performance G was unrelated to number of music courses and unrelated to number of physics courses, also, the background knowledge variables did not interact with type of text. Thus to the extent that the null hypothesis is supported, calibration of performance is unrelated to expertise.

The significant performance G is important in two respects. First, it replicates our previous finding (Glenberg & Epstein, 1985), and creates a bridge between our work on calibration of comprehension and other work on calibration of probabilities. The ability to accurately postdict performance has been a stable feature of the calibration literature (Lichtenstein et al., 1982).

Second, the significant perfomance G helps to rule out some uninteresting interpretations of the non-significant calibration of comprehension G. In particular, given that performance G is significant, it is less likely that the non-signifcant calibration of comprehension G reflects low statistical power, or any hidden constraints in our procedures.

Recalibration and Its Components

Insert Table 4 about here

Recalibration confidence (probe 4), n = 61. After assessing confidence in performance, subjects were asked for confidence in ability to answer a second inference test related to the same principle. Recalibration confidence is markedly similar to calibration confidence (probe 1). The recalibration confidence means were 4.67 (.87) and 4.72 (.88) for the music and physics texts respectively. The only significant effect was the interaction of text type and background knowledge, $F(4, 116) = 77.14$, MSE = .0022. The regression coefficients are given in Table 4. Note that for both variables, the difference between the coefficients for the music and physic texts is almost as great for recalibration confidence as for calibration confidence (Table 2).

Recalibration proportion correct (probe 5), n = 61. Performance on the second inference test was similar to performance on the first. The mean proportions correct were .73 (.13) and .79 (.12) for the music and physics texts, respectively. The difference was significant, $F(1, 116) = 21.48$, MSE = .0030.

There was also a significant interaction between type of text and background knowledge, $F(2, 116) = 10.61$, MSE = .0030. The regression coefficients are listed in Table 4. The only significant component in the interaction involves the number of physics courses variable. Increments in number of physics courses are associated with increments in proportion correct for the physics texts, but not for the music texts (this effect was not significant in the first analysis using four variables to code background knowledge).

As in the analysis of the first inference test, there was a main effect for dualism, $F(1, 55) = 8.15$, MSE = .0135, in the first set of analyses. On the average, a unit increase in the dualism variable was associated with a decrease of .0365 in proportion correct.

Recalibration G, n = 54. Recalibration Gs were .06 (.53) and .02 (.62) for the music and physics texts respectively. Neither was significantly different from zero. Background knowledge did account for a significant proportion of the variance in recalibration G, $F(2, 51) = 4.49$, MSE = .0167. Number of music courses was the variable that contributed most.

There was also a significant interaction between type of text and background knowledge, $F(2, 102) = 6.12$, MSE = .0032, that was carried by the physics courses variable. The regression coeficients for this interaction are in Table 4. As with initial calibration, increments in physics courses had a greater detrimental effect on recalibration for the physics texts than for the music texts.

The recalibration data do not replicate the effect reported by Glenberg and Epstein (1985). They found that recalibration was significantly greater than initial calibration (based on probes 1 and 2). Here, overall recalibration is not different from zero, and any effect of expertise is to decrease recalibration, much as it decreases initial calibration. This failure to replicate is addressed in the discussion.

Stability of Calibration Over Days, n = 61

Two new calibration Gs were computed for each subject, one for day 1 and one for day 2 of the experiment. Each of these Gs was based on probes 1 (initial confidence) and 2 (initial inference evaluation) for 16 texts, 8 music texts and 8 physics texts. All previously reported Gs were computed separately for different types of texts.

The across-text-type Gs were .18 (.54) and .30 (.45) for day 1 and day 2, respectively. Both of these Gs are significantly greater than zero, ts = 2.60 and 5.21, respectively.

The correlation between across-text-type G for day 1 and across-text-type G for day 2 was only -.03. This may be compared with the correlation between confidence (probe 1) on day 1 and day 2, .84, and the correlation between proportion correct on the two days, .37. This failure to find stable individual differences suggests that the search for variables (e.g., dualism) that would correlate with calibration is futile.

These data present somewhat of a mystery. Why should G computed by collapsing across type of text be significantly greater than zero, when calibration (based on the same number of texts) computed within a type of text is essentially zero? One rather uninteresting explanation is that G based on a single type of text suffers from a restricted range; combining across text types pools texts that have a greater range on both the confidence scale and proportion correct resulting in a larger G.

Two arguments can be made against this explanation. First, G, unlike the product-moment correlations requires only ordinal data. In fact, the value of the statistic is completely unaffected by the range of confidence scores, as long as there is some variability so that the statistic can be computed.

Second, recall that performance Gs were significantly greater than zero. These performance Gs use exactly the same proportion correct data as the calibration Gs that are not significantly different from zero. Clearly, the poor calibration Gs cannot be attributed to restricted range of performance.

A second explanation for the significant across-text-type Gs is provided by the following hypothesis. We suppose that subjects can

accurately classify themselves as relatively more expert in music or in physics. We also suppose that self-classified music students believe that they will do better on music texts than on physics texts, and that self-classifed physics students believe the opposite. In fact, these beliefs are consonant with the results of our analyses of proportion correct. Finally, we suppose that confidence is based on these beliefs. Because performance is better in texts in the domain consonant with the self-classification than in the other domain, the self-classification is indeed predictive of performance so that across-text-type $\underline{G}$ is greater than zero. According to this hypothesis, calibration across domains simply reflects the expert's use of base rates to accurately predict differences in performance across domains.

There is strong evidence consistent with the self-classification hypothesis. According to the hypothesis, subjects use their experience with music or physics to generate a confidence assessment for each text. This experience is public data, at least to the extent it is revealed on the questionnaire filled out at the end of the experiment (see Method section and Table 1). If the hypothesis is correct, we should be able to use these public data to generate confidence ratings that predict performance as well as the confidence ratings actually given by the subjects.

The test of this prediction required several steps. (A total of 43 subjects contributed to all steps.) First, a calibration $\underline{G}$ was computed for each subject using all 32 texts (to provide a maximally sensitive test). The average $\underline{G}$ was .20 (.35), which is significantly greater than zero, $\underline{t} = 3.75$. Next, using the regression coefficients for confidence listed in Table 2, we computed for each subject a single simulated confidence rating for music texts and a single simulated confidence rating for physics texts. Finally, using

these simulated confidence ratings a simulated G was computed for each subject.

The mean simulated G was .22 (.44). This G was significantly greater than zero, $t = 3.28$. The mean simulated G and the mean of the actual Gs (based on 32 texts) were not signifcantly different. Importantly, the correlation between the simulated Gs based on public data and the Gs based on the subjects' own 32 confidence ratings was .57.

An implication of the self-classification hypothesis is that subjects are not using any sort of privileged access to their own knowledge to generate confidence assessments; indeed the hypothesis implies that subjects are not assessing comprehension of the texts when they provide a confidence judgement, instead they are simply recording a belief based on their general experience. Thus the significant across-text-type G should not be taken as evidence of accurate self-assessments comprehension. As just demonstrated, the confidence scores generated by the regression equation, which obviously has no privileged access to subject's degree of comprehension, can predict performance as well as the subject's own confidence ratings.

A similar explanation can be applied to the significant correlation between average confidence and average performance. On day 1, the correlation was .51, and on day 2 the correlation was .37. These correlations do not imply that subjects are calibrated. Some subjects know that they generally do well on tests and hence have high confidence, other subjects know that they generally do poorly on tests and hence have low confidence. To the extent that past experience predicts future performance, there is a correlation between average confidence and performance. However, neither the subjects who generally do well nor those who generally do poorly can accurately assess comprehension and

predict which inference tests will be answered correctly: When calibration must be based on actual assessments of comprehension (i.e., within a text type) calibration is zero.

Discussion

This experiment was designed to answer four questions. The first question was whether calibration of comprehension for texts in a given domain changes with expertise in that domain. The answer is yes, but perhaps in an unexpected way. The regression analyses for both calibration and recalibration indicate that G decreases with experience in a domain (and significantly so for physics).

The second question was whether there are stable individual differences in calibration of comprehension. Here the answer is no. Even the significant across-text-type G was not stable across days.

The third question was whether accurate calibration of performance would be found. For this question the answer is yes. Calibration of performance was not only statistically significant, it was quite large, .42 for the music texts and .36 for the physics texts (recall that G is the difference between two probabilities). Apparently, subjects can fairly accurately judge the quality of their performance on an inference verification test.

The fourth question concerned recalibration. Previous results indicated that subjects could take advantage of experience gained while answering an inference test to predict performance on future tests over the same material. The subjects participating in this experiment did not exhibit this ability.

Self-classification Hypothesis

The pattern of the results discussed so far, as well as other data, is consistent with the self-classification hypothesis. The hypothesis is that

subjects classified themselves as relatively expert in music or physics, and used the belief that expertise in a domain is correlated with comprehension of texts in that domain to generate confidence ratings. That is, self-classification rather than assessment of text comprehension controlled the confidence ratings.

The strongest evidence consistent with the hypothesis is from the analysis of the simulated Gs. The mean simulated G was not significantly different from the mean G produced by the subjects, and the correlation between the simulated Gs and the actual across-text-type Gs was substantial.

The self-classification hypothesis provides a simple explanation for the poor calibration within a text type. According to the hypothesis, subjects are not actually assessing comprehension, instead they are responding on the basis of beliefs about their abilities within a given domain. These beliefs are not sufficiently fine-grained (differentiated) to accurately predict performance within a domain.

Variability of confidence ratings within a domain may be based on judged familiarity with a topic. In fact, the average correlation between familiarity ratings (obtained at the end of the second session) and confidence was .63 (.17). When these familiarity ratings (one for each text) are used to compute a G, the average familiarity G, .23 (.29), is not significantly different from the average simulated G based on a single confidence rating for each type of text. Thus, although the familiarity ratings account for varibility in the confidence ratings, they do not contain any useful information for predicting performance over and above that provided by the self-classifications.

The self-classification hypothesis is also at least partially consistent with the negative relationship between expertise and calibration (within a

domain). Most likely, only subjects who regard themselves as having some expertise will apply the self-classification strategy. Other subjects may actually carry out some form of evaluation of comprehension that predicts performance on the inference test (based on the regression equations, subjects with an average number of music courses, but no physics courses, were calibrated). Thus increasing expertise is associated with application of a less successful strategy for predicting performance within a domain.

The self-classification strategy was probably also applied when subjects were asked to re-assess confidence (probe 4) in future performance. The pattern of regression coefficients relating background knowledge to initial confidence (probe 1) was similar to the pattern relating background knowledge to re-assessed confidence (probe 4, compare Tables 2 and 4). Apparently subjects were using the same information (self-classifications) to make both ratings.

On the other hand, it appears that confidence in performance (probe 3) was not determined by self-classification. First, these confidence ratings were significantly correlated with actual performance (performance G greater than zero) within a domain of knowledge, which is not possible by application of the self-classification strategy alone. Second, the pattern of regression coefficients relating background knowledge to confidence in performance is quite different from the pattern relating background knowledge to initial confidence (compare coefficients in Table 3 to those in Table 2).

When is the Self-classification Strategy Applied?

We have stressed the contribution that self-classification may make to the computation of confidence. But we do not intend to imply that the metacognitive rule expressing the relationship between self-classification and likelihood of successful performance is the only rule for computing confidence. Other rules

based on familiarity and ease or completeness of access to the relevant text may also be engaged. In fact, earlier we reported a significant correlation between familiarity ratings and confidence ratings.

Given that there is a repertoire of metacognitive rules for computing confidence, when is the self-classification strategy applied? One consideration may be the task setting. Various aspects of the setting of the current experiment probably encouraged use of the strategy. Subjects knew that they were selected on the basis of their experience in music and physics courses. In addition, the texts were clearly in one domain or the other, and the contrast was heightened by the presentation order which alternated texts from the two domains. Probably, the strategy is encouraged whenever the domain of the text clearly matches the subject's own beliefs about domains of expertise.

In addition to the task setting, it is plausible to postulate that other factors affecting availability of rules in memory are involved in determining the subject's choice from the repertoire of metacognitive rules. Also, it seems likely that the process of selection is dynamic reflecting the effects of several variables operating concurrently to assign prominence to different metacognitive rules. The dynamic character of the process helps us to formulate a coherent account of the principal findings of this study.

We have argued that the initial confidence rating was computed by application of the self-classification strategy, the rule made most available by the task setting. Why then, was the self-classification strategy not applied when rating confidence in performance? After answering the first inference test (probe 2), subjects could base their confidence rating on either the self-classification strategy, or the specific experience gained from answering the inference (such as ease of retrieving relevant propositions from memory).

We propose that most subjects chose to use specific experience for the following reasons. (a) Having just evaluated the inference (probe 2), the experience was probably highly available while making the confidence in performance rating (probe 3). (b) Some of the specific experiences were probably easily recognized as diagnostic. For example, failure to retrieve any information relevant to evaluating the inference is easily recognized as a useful predictor of chance performance. (c) The experience was specific to the particular judgement being made, whereas the self-classification strategy is more general. Thus after answering the first inference other metacognitive rules (e.g., base confidence on experience, perhaps latency, answering the question) are at least as available as the self-classification strategy.

On the other hand, it appears that the self-classification strategy was applied again in generating predictions about future performance on the recalibration confidence rating (probe 4, see discussion of recalibration). Why do subjects revert to using the self-classification strategy for probe 4, after rejecting it for probe 3? In answering probe 4, subjects also have a choice of metacognitive rules. We suspect that the self-classification strategy is chosen because of a difference in the diagnostic value attributed by the subject to the experience gained from answering the initial inference. Experience answering the first inference is believed to be diagnostic for judging performance on the first inference. The experience is believed to have less diagnostic value for predicting future performance. Given the belief that the diagnostic value of the experience is low and the ready availability of a strategy with high face validity, subjects chose the self-classification strategy.

Use of the self-classification strategy when answering probe 4 helps to explain why significant recalibration was not found in this experiment, but was

found in Glenberg and Epstein (1985). As discussed before, the self-classification strategy cannot produce calibration within a domain, obviating any possibility of significant recalibration. In Glenberg and Epstein (1985) the texts were sampled from a variety of domains, reducing availability and use of the self-classification strategy. Thus in our previous research, when subjects re-assessed confidence after the initial inference test, it is likely that the subjects were forced to use a metacognitive role with greater predictive validity than the self-classification strategy.

In summary, it appears that the self-classification strategy will be used (and be effective) under the following conditions. First, the structure of the calibration task suggests the strategy by highlighting the relationship between a reader's domain of knowledge and the domain of the text. Second, the reader does not have available information that is believed to be more specific or more diagnostic than self-classification. Whether or not application of the strategy produces calibration depends at least in part on the structure of the task. Application of the strategy across domains of expertise is almost guaranteed to produce high calibration. Unfortunately, the self-classification strategy alone cannot produce calibration within a domain of expertise.

References

Birkmire, D.P. Effect of the interaction of text structure, background knowledge, and purpose on attention to text. (Technical memorandum 6-82) U.S. Army Human Engineering Laboratory, Aberdeen Proving Ground, Maryland.

Bradley, J. V. (1981). Overconfidence in ignorant experts. Bulletin of the Psychonomic Society, 17, 82-84.

Glenberg, A. M., & Epstein, W. (1985). Calibration of Comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 702-718.

Hock, S.J. (1985). Counterfactual reasoning and accuracy in predicting personal events. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 719-731.

Kassin, S.M. (1985). Eyewitnesss identification: Retrospective self-awareness and the accuracy-confidence correlation. Journal of Personality and Social Psychology, 49, 878-893.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.) Judgement under certainty: Heuristics and biases. New York: Cambridge University Press.

Nelson, T.O. (1984). A comparison of current measures of the accuracy of feeling of knowing predictions. Psychological Bulletin, 95, 109-133.

Oskamp, S. (1965). Overconfidence in case study judgements. Journal of Consulting and Clinical Psychology, 29, 261-265.

Ryan, M. P. (1984). Monitoring text comprehension: Individual differences in epistemological standards. Journal of Educational Psychology, 76, 248-258.

Author Notes

This research was funded by Office of Naval Research Contract N0014-85-K-0644 and National Institute of Education Grant NIE-G081-0009 to the Wisconsin Center for Education Research. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Institute of Education or the Department of Education.

Our thanks to Craig Morris, Tom Sanocki, and Naomi Swanson for assisting in execution of this research.

Requests for reprints should be sent to either author at the Department of Psychology, W. J. Brogden Psychology Building, University of Wisconsin, Madison, Wisconsin 53706.

Table 1

Subject Characteristics

Variable	Mean	SD	Smallest	Largest
Dualism	2.59	0.80	1.14	4.14
Music courses	2.76	3.77	0.00	15.00
Music experience	1.34	0.96	0.00	3.00
Physics courses	2.56	2.38	0.00	11.00
Physics experience	0.26	0.49	0.00	2.00

Table 2

Regression Coefficients for Calibration and Its Components

Dependent variable	Independent Variable: Y-Intercept.	Music Courses	Physics Courses
Music text confidence	4.7471	0.1003a	-0.1300b
Physics text confidence	4.5301	-0.0789a	0.1601b
Music text prop. cor.	0.6453c	0.0121d	*0.0159
Physics text prop. cor.	0.7275c	-0.0022d	*0.0275
Music text G	0.1034	-0.0251	0.0120e
Physics text G	0.3740	-0.0213	-0.1165e

Note: Asterisks indicate the coefficients of variables having significant main effects (significantly related to the dependent variable averaged over text type). Coefficients with the same letter are significantly different from one another and indicate a significant interaction between the independent variable and text type.

Table 3

Regression Coefficients for Performance Confidence and Calibration

Dependent variable	Independent Variable: Y-Intercept.	Music Courses	Physics Courses
Music text confidence	4.7179a	0.0775b	-0.0671c
Physics text Confidence	4.8523a	-0.0377b	0.0910c
Average G	0.4517	-0.0081	-0.0154

Note: Coefficients with the same letter are significantly different from one another and indicate a significant interaction between the independent variable and text type.

Table 4

Regression Coefficients for Recalibration and Its Components

Dependent variable	Independent Variable: Y-Intercept.	Music Courses	Physics Courses
Music text confidence	4.6512	0.0944a	-0.0961b
Physics text confidence	4.5287	-0.0667a	0.1421b
Music text prop. cor.	0.7048c	0.0060	0.0012d
Physics text prop. cor.	0.7301c	0.0000	0.0224d
Music text G	-0.0596	*0.0309	0.0098e
Physics text G	0.1768	*0.0277	-0.0918e

Note: Asterisks indicate the coefficients of variables having significant main effects (significantly related to the dependent variable averaged over text type). Coefficients with the same letter are significantly different from one another and indicate a significant interaction between the independent variable and text type.

Appendix

Organic Unity - Text

The way in which the parts of a musical work relate to form a whole has long been an important consideration of musical aesthetics. The theory of organic unity, which directly compared the parts and whole of musical works to those of living things, became part of the evaluative process as an aesthetic norm in the early 19th century. According to the theory, musical pieces were analogous to creatures: Each part of a successful work was essential, just as every part of the body was (supposedly) essential; no part of a good piece of music could be substituted for another, since each had a specific function in the unified whole. Furthermore, as in an organic body, the combined functions of all the parts of a musical masterwork were believed to form a coherent unity because of specific relationships which held the parts together; thus no part of the whole could stand separately as a successful work. Certain parts of the whole were believed to carry more important functions than others, just as the heart has a more important function than the little toe. Furthermore, it was believed that great composers were great creators, who, like God, fashioned "living organisms." (Consider a statement by Karl Kahlert, music aesthetician, writing in 1848: "What is musical form but the natural body that music must assume in order to establish itself as a living organism?"). Though the analogy is useful and interesting, problems with the theory of organic unity are evident. It assumed that composers were aiming at a particular kind of structural unity, which was simply not the case for most pieces written before about 1600 or after about 1910. It demonstrated an evaluative bias against longer forms, especially opera, where the semblance of complete unity was more difficult to maintain.

Probe 1 - Confidence Scale

Organic Unity

Circle a single number on the following scale to report your confidence in being able to accurately judge the correctness of an inference drawn from the reading about the relationships between parts of a composition according to the theory of organic unity.

Probe 2 - Initial Inference

Organic Unity

Inference: According to the theory of organic unity, it is not possible to improve some compositions by deleting specific parts.

T F

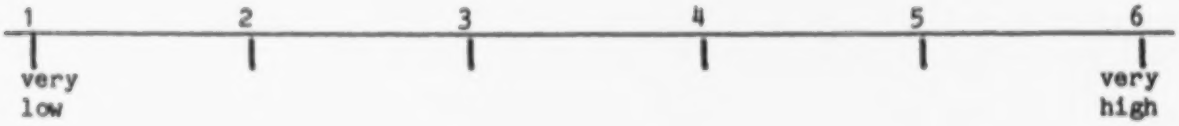
Phase 3 - Confidence in Performance

Organic Unity

Circle a single number on the following scale to report your confidence that you have answered the inference correctly.

Probe 4 - Recalibration Confidence

Circle a single number on the following scale to report your confidence that you can judge the correctness of another inference drawn from the reading about the relationships between parts of a composition according to the theory of organic unity.



Probe 5 - Second Inference

Organic Unity

Inference: The theory of organic unity does not explain why a single movement of a work is often complete and performable without the other movements of the composition.

T F